DOCUMENT RESUME

ED 282 921 TM 870 339

AUTHOR Cope, Ronald T.
TITLE A Generalizability Study of the Angoff Method Applied to Setting Cutoff Scores of Professional Certification Tests.
PUB DATE Apr 87
NOTE 16p.; Paper presented at the Annual Meeting of the American Educational Research Association (Washington, DC, April 20-24, 1987).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Certification; Correlation; *Cutting Scores; Error of Measurement; *Feedback; *Generalizability Theory; *Interrater Reliability; *Judges; Licensing Examinations (Professions); Statistical Analysis; *Test Items; Validity
IDENTIFIERS *Angoff Methods; Standard Setting

ABSTRACT

This study used generalizability theory and other statistical concepts to assess the application of the Angoff method to setting cutoff scores on two professional certification tests. A panel of ten judges gave pre- and post-feedback Angoff probability ratings of items of two forms of a professional certification test, and another panel of nine judges similarly rated and received feedback for the items of two forms of a similar test for more advanced certification. Feedback for a new item consisted of mean judge rating of the item. Feedback for an old item consisted of mean rating plus the item's p-value for a group of examinees who had scored close to the previous cutoff score. Results show that: (1) mean post-feedback ratings were significantly greater than initial ratings for all four test forms; (2) hypothesized greater internal consistency of second ratings was observed only for the first judge group and test; (3) correlations of mean (over judges) first and second ratings were lower for old items than for new items; (4) judges varied widely in the correlation of their first and second ratings; and (5) most judges were strongly influenced by the marginal-group p-values given for old items. (Author/JAZ)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED282921

A Generalizability Study of the Angoff Method Applied to Setting Cutoff Scores of Professional Certification Tests

Ronald T. Cope

The American College Testing Program

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

R.T. Cope

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Paper presented at the Annual Meeting of the American Educational Research Association

Washington, D.C., April, 1987

TM 870 339

BEST COPY AVAILABLE

Abstract

A panel of ten judges gave pre- and post-feedback Angoff probability ratings of items of two forms of a professional certification test, and another panel of nine judges similarly rated and received feedback for the items of two forms of a similar test for more advanced certification. Feedback for a new item consisted of mean judge rating of the item. Feedback for an old item consisted of mean rating plus the item's p-value for a group of examinees who had scored close to the previous cutoff score.

Results show that (1) mean post-feedback ratings were significantly ($p < .001$) greater than initial ratings for all four test forms, (2) hypothesized greater internal consistency of second ratings was observed only for the first judge group and test, (3) correlations of mean (over judges) first and second ratings were lower for old items than for new items, (4) judges varied widely in the correlation of their first and second ratings, (5) most judges were strongly influenced by the marginal-group p-values given for old items.

A Generalizability Study of the Angoff Method Applied
to Setting Cutoff Scores of Professional Certification Tests

The purpose of this study is to use generalizability theory and other statistical concepts to assess the application of the Angoff method (Angoff, 1971, pp. 514-515) to setting cutoff scores on two professional certification tests. The Angoff method requires each member of a panel of judges to independently estimate the probability that a hypothetical "minimally qualified" examinee will correctly answer a given item of a test. These estimated probabilities are obtained for each item of the test for which one wishes to obtain a cutoff score. Next each rater's estimated probabilities are added to obtain judge-estimated cutoff scores. Finally, these cutoff scores are averaged across judges. The result is the cutoff score.

Generalizability theory (Cronbach, Gleser, Nanda, and Rajaratnam, 1972; Brennan, 1983) can be used to estimate variance components of sources of error in judgmental standard setting (Brennan and Lockwood, 1980; Kane and Wilson, 1984) under the assumption that judges and items constitute random samples from populations of judges and items.

One might expect that giving normative feedback to judges would have the effect of increasing the internal consistency of ratings: judges giving aberrant ratings would adjust their ratings to be closer to the mean rating. Accordingly, it was hypothesized that second ratings would show greater internal consistency than initial ratings.

Method

A panel of ten judges gave Angoff estimates of two forms of a professional certification test. A panel of nine other judges rated items of two forms of a similar test used for further certification. For each item, each judge supplied two estimates of the Angoff probability—an initial rating

and a post-feedback rating. After their first ratings of a test section's items, judges were given the following kinds of feedback: (1) for old or previously used items, proportion ("p-value") of a group of examinees who scored close to a previous cutoff score and correctly answered the item, plus the means of the judges' ratings of each item, (2) for new items, mean judge ratings only. Judges were told to regard the feedback as general reference, not as targets to shoot at. They were also told to take into consideration chance successes on the items. This procedure incorporates some features of the one used by Jaeger (1982).

The generalizability analysis used a judge by items-nested-within subject-category design. Item-level observations (judge ratings of items) are given by the model,

$$X_{jic} = \mu_c + \kappa_j + \lambda_i + \pi_{jc} + \beta_{i:c} + \pi\beta_{ji:c} ,$$

where X_{jic} represents the rating of the jth judge to the ith item within category c. The μ_c constitute fixed effects and the remaining terms are random effects having means of zero. This design is unbalanced owing to different numbers of items in different categories. But because the number of items in a given category is fixed, category constitutes a fixed facet of observation, and the problem of estimating variance components in unbalanced designs presents no great difficulty here.

For each form, variance components were estimated separately for sets of first and second ratings. To test the hypothesis of differences of internal consistency between first and second ratings, two reliability-like coefficients were computed for each set of ratings: generalizability coefficient (Brennan, 1983, p. 17) and phi (Brennan, 1983, p. 108).

Additional statistics obtained for all sets of ratings and for total items, old items, and new items include means and standard deviations, correlation of first and second ratings for each judge and for mean ratings over judges, and correlation of first and second ratings with marginal group p-value (old items only). t-tests were computed to test for significance of differences between mean (over judges) first and second item ratings.

Results

Differences Between First and Second Ratings

In Table 1 the two-tailed t-tests for dependent samples show significantly greater ($p < .001$) mean second ratings than first ratings for all forms, and for old-, new-, and total items.

Insert Table 1 about here

Variance Components and Internal Consistency of Ratings

Table 2 lists the following estimated variance components: judge, item subject category, judge-by-category interaction, item within category, and judge-by-item interaction within category. First and second ratings show similar relative magnitudes of most variance components, but second rating components tend to be smaller, especially for item category. Also, category components of second ratings tend to be about half those of first ratings. Except for second ratings of Form 2A items, judge-by-item within category yielded the largest variance components.

Insert Table 2 about here

Table 3 lists the generalizability coefficient and phi for first and second item ratings. Increases of reliability-like coefficients for Forms 1A and 1B were consistent with the hypothesized increase of internal consistency from first to second ratings. However, these coefficients stay the same or decrease slightly for the two forms rated by the second group of judges.

Insert Table 3 about here

Effect of Feedback to Raters: Intercorrelations of first rating, second rating, and marginal-group p-values

The correlations shown in Table 4 provide a way of assessing the effect of feedback on judge ratings. First and second ratings of new items correlate quite highly for all four forms, from .93 to .95. In contrast, old items--for which marginal group p-values were given after first ratings--show notably lower first rating-second rating correlations for forms 1B, 2A, and 2B. Not surprisingly, marginal group p-values correlate higher with second ratings than with first ratings.

Insert Table 4 about here

Table 4 presents evidence of a difference in the two judge groups' uses of the marginal group p-values: Correlations of marginal group p-value with second rating are higher for judge group 2 (.83 and .72) than for group 1 (.40 and .53). In accord with these results, the first rating-second rating correlations for old items are higher for group 1 (.90 and .80) than for group 2 (.68 and .69), even though such correlations for new items are virtually the same for both groups (.93 and .95 vs. .94 and .93).

Table 5 presents correlations of first and second ratings by judge. They range from nearly perfect (Form 1A, judge 4) down to .37 (Form 2B, judge 4). In agreement with the Table 4 correlations of ratings averaged over judges, Group 1 judges tended to have higher correlations of first and second ratings than did Group 2 judges, and correlations for old items tended to be lower than those for new items. However, some judges (most of them in Group 1) had higher correlations for old items than for new items.

Insert Table 5 about here

Judges tended to be quite consistent across forms in their degree of first rating-second rating consistency: Correlating the Form A-Form B correlations yielded .79 (old-old) and .93 (new-new) for Group 1 and .96 (old-old) and .89 (new-new) for group 2.

Discussion

To feed back or not to feed back? Is it better to take a judge's pristine first ratings unadulterated by knowledge of other judges' ratings or of past performance standards? If judges are given marginal-group p-values, and if they take these numbers very seriously--regard them as targets, as criteria--then the cut score study becomes a charade: one expects virtual duplication of old standards, correct or not. Yet in the absence of some kind of reality check, panels of judges have occasionally created havoc by arriving at ridiculously high or low standards.

The results of this study give no universal answer to this question of first ratings only versus feedback and second rating. But they do confirm that giving feedback makes a difference, and that consequently the decision

whether to use feedback is an important consideration in a judgmental cut score study.

Although not specifically addressed in this study, the matter of training and instructing judges merits investigation. Seemingly trivial differences in instructions might affect outcomes greatly. There is an indication of this effect in the differences between correlations of first and second ratings for the two groups of judges. The intent was to give identical instructions to both groups. However, the lower correlations for the second group of nine suggest that judges might have somehow taken more seriously or given more weight to or employed differently the marginal p-values than did members of the other group. This is something to worry about. Can we ever determine the exactly "right" instructions that yield the "right" cut score?

Validity-related considerations in judgmental standard setting are turning out to be much more vexing than those of reliability. Generalizability concepts and techniques furnish convenient tools for assessing and if need be increasing the stability of estimates. One can sample more judges, perhaps also have the judges rate on multiple occasions (with or without feedback), and thereby obtain very stable estimates. But what is being estimated? So much depends upon judges' collective grasp of elusive notions like "minimal competence." Explicating such notions and using them to arrive at the "true" cut score—or something very close to it—poses a challenge to contemporary standard setters.

References

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), Educational Measurement (2nd ed.). Washington, D.C.: American Council on Education, 508-600.

Brennan, R. L. (1983). Elements of generalizability theory. Iowa City, IA: American College Testing Program.

Brennan, R. L., & Lockwood, R. E. (1980). A comparison of the Nedelsky and Angoff cutting score procedures using generalizability theory. Applied Psychological Measurement, 4, 219-240.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: Wiley.

Jaeger, R. M. (1982). An iterative structured judgment process for establishing standards on competency tests: Theory and application. Educational Evaluation and Policy Analysis, 4, 461-475.

Kane, M., & Wilson, J. (1984). Errors of measurement and standard setting in mastery testing. Applied Psychological Measurement, 8, 107-115.

Acknowledgments

I thank Dr. H. A. Wilson for collecting and supplying the data used in this study, Dr. Michael Kolen and Mr. Jerry Traeger for assistance in managing the data base, and Dr. Robert Brennan, Dr. Michael Kane, and Dr. Michael Kolen for comments on drafts of this paper.

Table 1

Differences between First and Second Item Ratings

Judge Group	Test Form	Number of Items	First Rating Mean	First Rating SD	Second Rating Mean	Second Rating SD	t	df	p
			All Items						
1	1A	147	.6616	.101	.6894	.087	-8.24	146	.000
	1B	149	.6811	.109	.7100	.098	-6.18	148	.000
2	2A	150	.6543	.098	.7193	.098	-10.94	149	.000
	2B	149	.6598	.095	.7188	.093	-10.59	148	.000
			Old Items						
1	1A	97	.6756	.096	.7003	.084	-5.92	96	.000
	1B	95	.6867	.107	.7167	.099	-4.43	94	.000
2	2A	111	.6613	.100	.7351	.098	-9.62	110	.000
	2B	112	.6688	.099	.7337	.093	-9.06	111	.000
			New Items						
1	1A	50	.6344	.106	.6682	.089	-5.94	49	.000
	1B	54	.6711	.113	.6981	.096	-5.33	53	.000
2	2A	39	.6344	.089	.6741	.086	-8.19	38	.000
	2B	37	.6324	.080	.6735	.075	-8.58	36	.000

Note. There were ten judges in Group 1 and nine in Group 2; however, Judge 10 was dropped from the Form 1B calculations because of missing first ratings of 11 category C items.

Table 2

Variance Components of Item Ratings:
Judges by Items-Nested-Within-Content-Categories Design

Form	Component	df	First Rating			Second Rating		
			MS[a]	Variance Component	SEV[b]	MS[a]	Variance Component	SEV[b]
	J	9	.35030	.00228	.00102	.43771	.00290	.00127
	C	7	.71628	.00340	.00184	.38971	.00170	.00100
1A	JC	63	.03589	.00109	.00034	.02970	.00098	.00028
	I:C	139	.07093	.00552	.00085	.06012	.00484	.00072
	JI:C	1251	.01573	.01573	.00063	.01169	.01169	.00047
	J	8	.30355	.00194	.00091	.39488	.00257	.00119
	C	7	.64183	.00323	.00181	.37751	.00175	.00106
1B	JC	56	.03501	.00109	.00035	.02436	.00068	.00024
(Judge	I:C	141	.07984	.00724	.00105	.07185	.00668	.00095
10	JI:C	1128	.01472	.01472	.00062	.01171	.01171	.00049
omitted)								
	J	8	.54883	.00358	.00164	.36199	.00236	.00108
	C	6	.52470	.00217	.00136	.31993	.00118	.00083
2A	JC	48	.04981	.00174	.00047	.02223	.00069	.00021
	I:C	143	.06807	.00617	.00089	.07716	.00776	.00101
	JI:C	1144	.01251	.01251	.00052	.00735	.00735	.00031
	J	8	.56229	.00369	.00169	.43599	.00287	.00131
	C	6	.48779	.00200	.00127	.29091	.00105	.00076
2B	JC	48	.05171	.00183	.00049	.02951	.00101	.00028
	I:C	142	.06490	.00580	.00085	.06730	.00659	.00088
	JI:C	1136	.01274	.01274	.00053	.00799	.00799	.00034

[a]MS--mean square
[b]SEV--standard error of variance component

Table 3

Reliability-like Measures of Internal Consistency:
Generalizability Coefficient, Phi

Form	Judge Group	Generalizability Coefficient		Phi	
		Rating 1	Rating 2	Rating 1	Rating 2
1A	1	.96	.97	.94	.96
1B	1	.95	.97	.93	.95
2A	2	.98	.98	.97	.96
2B	2	.98	.98	.97	.97

Table 4

Correlations of Mean First Rating, Mean Second Rating, and Marginal Group p-Value

Judge Group	Test Form	Number of Items	Correlation: Rating 1/ Rating 2	Correlation: Rating 1/ p-Value	Correlation: Rating 2/ p-Value
			Old Items		
1	1A	97	.90	.21	.40
	1B	95	.80	.32	.53
2	2A	111	.68	.30	.83
	2B	112	.69	.20	.72
			New Items		
1	1A	50	.93	No p-values available for new items.	
	1B	54	.95		
2	2A	39	.94		
	2B	37	.93		

Table 5

Correlation of Judges' First and Second Item Ratings

Judge Group	Judge	Test	Form A		Form B	
			Old Items	New Items	Old Items	New Items
1	1	1	.74	.61	.73	.65
	2		.95	.95	.90	.96
	3		.87	.85	.84	.84
	4		1.00	.99	.90	1.00
	5		.82	.90	.84	.91
	6		.78	.93	.64	.84
	7		.87	.91	.83	.98
	8		.91	.90	.81	.93
	9		.83	.92	.71	.96
	10		.74	.67	.75	.72
2	1	2	.40	.64	.43	.62
	2		.52	.72	.52	.71
	3		.67	.91	.76	.95
	4		.41	.76	.37	.81
	5		.68	.80	.70	.79
	6		.49	.78	.40	.63
	7		.81	.92	.94	.94
	8		.77	.78	.79	.82
	9		.57	.65	.66	.50